

Word2vecを用いた単語ベクトル算出

- ①「青空文庫」より夏目漱石の「三四郎」をダウンロードして
「sanshiro.tex」というファイル名でpythonのエディタ例えばidleが
あるフォルダに保存する
- ②コマンドプロンプトあるいはWindows PowerShell を用いて次の
コマンドでpythonのライブラリーをダウンロード
 pip install janome
 pip install gensim

青空文庫のテキストファイルの冒頭

三四郎
夏目漱石

【テキスト中に現れる記号について】

《 》：ルビ
(例) 頓狂《とんきょう》

|：ルビの付く文字列の始まりを特定する記号
(例) 福岡県|京都郡《みやこぐん》

[#]：入力者注 主に外字の説明や、傍点の位置の指定
(数字は、JIS X 0213の面区点番号またはUnicode、底本のページと行数)
(例) ※ [# 「魚+師のつくり」、第4水準2-93-37]

[]：アクセント分解された欧文をかこむ
(例) [ve'rite] 《ヴェリテ》 [vraie] 《ヴレイ》
アクセント分解についての詳細は下記URLを参照してください
http://www.aozora.gr.jp/accnt_separation.html

[# 7 字下げ] — [# 「一」は中見出し]

うとうととして目がさめると女はいつのまにか、隣のじいさんと話を始めている。このじいさんはたしかに前の前の駅から乗ったいなか者である。発車まぎわに頓狂《とんきょう》な声を出して駆け込んで来て、いきなり肌《はだ》をぬいだと思っただけの中にお灸《きゅう》のあとがいっぱいあったので、三四郎《さんしろう》の記憶に残っている。じいさんが汗をふいて、肌を入れて、女の隣に腰をかけたまですぐ注意して見ていたくらいである。

うとうととして目がさめると女はいつのまにか、隣のじいさんと話を始めている。このじいさんはたしかに前の前の駅から乗ったいなか者である。発車まぎわに頓狂な声を出して駆け込んで来て、いきなり肌を

取りかかる。与次郎だけが三四郎のそばへ来た。

「どうだ森の女は」
「森の女という題が悪い」
「じゃ、なんとすればよいんだ」
三四郎はなんとも答えなかった。ただ口の中で迷羊、迷羊と繰り返した。

「野々宮さんば、招待状を引き平切って床の上に捨てた。やがて先生とともにほかの絵の評に取りかかる。与次郎だけが三四郎のそばへ来た。」
「どうだ森の女は」
「森の女という題が悪い」
「じゃ、なんとすればよいんだ」
三四郎はなんとも答えなかった。ただ口の中で迷羊《ストレイ・シープ》、迷羊《ストレイ・シープ》と繰り返した。

底本：「三四郎」角川文庫クラシックス、角川書店
1951（昭和26）年10月20日初版発行
1997（平成9）年6月10日127刷
初出：「朝日新聞」
1908（明治41）年9月1日～12月29日

入力：古村充
校正：かとうかおり
2000年7月1日公開
2014年6月19日修正
青空文庫作成ファイル：
このファイルは、インターネットの図書館、青空文庫 (<http://www.aozora.gr.jp/>) で作られました。入力、校正、制作にあたったのは、ボランティアの皆さんです

```
import codecs
# ファイル読み込み、内部表現化
f = codecs.open('sanshiro.txt', "r", "sjis")
text = f.read()
f.close()
```

re 削除のコマンド

```
# ファイル整形
import re
# ヘッダ部分の除去
text = re.split('¥-[5,]', text)[2]
# フッタ部分の除去
text = re.split('底本:', text)[0]
# | の除去
text = text.replace('|', '')
# ルビの削除
text = re.sub('《.+?》', '', text)
# 入力注の削除
text = re.sub('[# .+?]', '', text)
# 空行の削除
text = re.sub('¥n¥n¥n', '¥n', text)
text = re.sub('¥¥r', '', text)
```

整形結果確認

```
# 頭の100文字の表示
print(text[:100])
# 見やすくするため、空行
print()
print()
# 後ろの100文字の表示
print(text[-100:])
```

実行結果

青空文庫のテキストファイルの末尾

Janomeを用いて単語のリストを作成

Janomeの中のTokenizerを用いて、名詞、動詞などに分ける関数を定義

三四郎
京都
ちょっと
用
ある
降りる
ついで
一
する
目
さめる
女
隣
じいさん
話
始める
いる

テスト結果

```
## Step2 Janomeを使い三四郎テキストから単語リストを生成

# Janomeのインストール
#!pip install janome

# Janomeのロード
from janome.tokenizer import Tokenizer

# Tokenizerインスタンスの生成
t = Tokenizer()

# テキストを引数として、形態素解析の結果、名詞・動詞原型のみを配列で抽出する関数
def extract_words(text):
    tokens = t.tokenize(text)
    return [token.base_form for token in tokens
            if token.part_of_speech.split(',')[0] in ['名詞', '動詞']]

# 関数テスト
ret = extract_words('三四郎は京都でちょっと用があって降りたついでに。')
for word in ret:
    print(word)

# 全体のテキストを句点('。')で区切った配列にする。
sentences = text.split('。')
# それぞれの文章を単語リストに変換(処理に数分かかります)
word_list = [extract_words(sentence) for sentence in sentences]

# 結果の一部を確認
for word in word_list[0]:
    print(word)
```

うとうととして目がさめると女はいつのまにか、隣のじいさんと話を始めている。このじいさんはたしかに前の前の駅から乗ったいなか者である。発車まぎわに頓狂《とんきょう》な声を出して駆け込んで来て、いきなり肌《はだ》をぬいだと思ったら背中にお灸《きゅう》のあとがいっぱいあったので、三四郎《さんしろう》の記憶に残っている。じいさんが汗をふいて、肌を入れて、女の隣に腰をかけたまでよく注意して見ていたくらいである。

女とは京都からの相乗りである。乗った時から三四郎の目についた。第一色が黒い。三四郎は九州から山陽線に移って、だんだん京大阪へ近づいて来るうちに、女の色が次第に白くなるのでいつのまにか故郷を遠のくような哀れを感じていた。それでこの女が車室にはいつて来た時は、なんとなく異性の味方を得た心持がした。この女の色はじっさい九州色《きゅうしゅういろ》であった。

三輪田《みわた》のお光《みつ》さんと同じ色である。国を立つまぎわまでは、お光さんは、うるさい女であった。そばを離れるのが大いにありがたかった。けれども、こうしてみると、お光さんのようなものもけって悪くはない。

ただ顔だちからいうと、この女のほうがよほど上等である。口に締まりがある。目がはっきりしている。額がお光さんのようにだだっ広くない。なんとなくいい心持ちにできあがっている。それで三四郎は五分に一度ぐらゐは目を上げて女の方を見ていた。時々女と自分の目がゆきあたることもあった。じいさんが女の隣へ腰をかけた時などは、もっとも注意して、できるだけ長いあいだ、女の様子を見ていた。その時女はにこりと笑って、さあおかけと言ってじいさんに席を譲っていた。それからしばらくして、三四郎は眠くなって寝てしまったのである。

その寝ているあいだに女とじいさんは懇意になって話を始めたものとみえる。目をあけた三四郎は黙って二人《ふたり》の話を聞いていた。女はこんなことを言う。――

子供の玩具《おもちゃ》はやっぱり広島より京都のほうが安くっていいものがある。京都でちょっと用があって降りたついでに、蛸薬師《たこやし》のそばで玩具を買って来た。久しぶりで国へ帰って子供に会うのはうれしい。しかし夫の任送りがとぎれて、しかたなしに親の里へ帰るのだから心配だ。夫は呉《くれ》にいて長らく海軍の職工をしていたが戦争中は旅順《りょじゅん》の方に行っていた。戦争が済んでからいったん帰って来た。まもなくあっちのほうが金がもうかるといって、また大連《たいれん》へ出かせぎに行った。はじめのうちは音信《たより》もあり、月々のものもちゃんちゃんと送ってきたからよかったが、この半年ばかり前から手紙も金もまるで来なくなってしまった。不実な性質《たち》ではないから、大丈夫《だいじょうぶ》だけれども、いつまでも遊んで食べているわけにはゆかないので、安否のわかるまではしかたがないから、里へ帰って待っているつもりだ。

じいさんは蛸薬師も知らず、玩具にも興味がないとみえて、はじめのうちはただはいはいと返事だけしていたが、旅順以後急に同情を催して、それは大いに気の毒だと言いだした。自分の子も戦争中兵隊にとられて、とうとうあっちで死んでしまった。いったい戦争はなんのためにするものかわからない。あとで景気でもよくなればだが、大事な子は殺される、物価《しよしき》は高くなる。こんなばかげたものはない。世のいい時分に出かせぎなどというものはなかった。みんな戦争のおかげだ。なにしろ信心《しんじん》が大切だ。生きて働いているに違いない。もう少し待ってればきつと帰って来る。――じいさんはこんな事を言って、しきりに女を慰めていた。やがて汽車がとまったら、ではお大事にと、女に挨拶《あいさつ》をして元氣よく出て行った。

じいさんに続いて降りた者が四人ほどあったが、入れ代って、乗ったのはたった一人《ひとり》しかない。もとから込み合った客車でもなかったのが、急に寂しくなった。日の暮れたせいかもしれない。駅夫が屋根をどしどし踏んで、上から灯《ひ》のついたランプをさしこんでゆく。三四郎は思い出したように前の停車場《ステーション》で買った弁当を食いだした。

車が動きだして二分もたったろうと思うころ、例の女はすうと立って三四郎の横を通り越して車室の外へ出て行った。この時女の帯の色がはじめて三四郎の目にはいった。三四郎は鮎《あゆ》の煮びたしの頭をくわえたまま女の後姿を見送っていた。便所に行ったんだなと思いながらしきりに食っている。

女はやがて帰って来た〔#「帰って来た」は底本では「帰った来た」〕。今度は正面が見えた。三四郎の弁当はもうしまいがけである。下を向いて一生懸命に箸《はし》を突っ込んで二口三口はおぼったが、女は、どうもまだ元の席へ帰らないらしい。もしやと思って、ひょいと目を上げて見るとやっぱり正面に立っていた。しかし三四郎が目を上げると同時に女は動きだした。ただ三四郎の横を通って、自分の座へ帰るべきところを、すぐと前へ来て、からだを横へ向けて、窓から首を出して、静かに外をながめだした。風が強くとあたって、髪《びん》がふわふわするところが三四郎の目にはいった。この時三四郎はからになった弁当の折《おり》を力いっばいに窓からほうり出した。女の窓と三四郎の窓は一軒おきの隣であった。風に逆らってなげた折の蓋《ふた》が白く舞いもどったように見えた時、三四郎はとんだことをしたのかと気がついて、ふと女の顔を見た。顔はあいにく列車の外に出ていた。けれども、女は静かに首を引っ込めて更紗《さらさ》のハンケチで額のところを丁寧にふき始めた。三四郎はともかくもあやまるほうが安全だと考えた。

「ごめんなさい」と言った。

女は「いいえ」と答えた。まだ顔をふいている。三四郎はしかたなしに黙ってしまった。女も黙ってしまった。そうしてまた首を窓から出した。三、四人の乗客は暗いランプの下で、みんな寝ぼけた顔をしている。口をきいている者はだれもない。汽車だけがすさまじい音をたてて行く。三四郎は目を眠った。

「しばらくすると『名古屋はもうじきでしょうか』と言う女の声をした。見るといつのまにか向き直って、及び腰になって、顔を三四郎のそばまでもって来ている。三四郎は驚いた。

「そうですね」と言ったが、はじめて東京へ行くんだからいっこう要領を得ない。

「この分では遅れますでしょうか」

「遅れるでしょう」

「あんたも名古屋へお降《お》りで……」

「はあ、降ります」

この汽車は名古屋どまりであった。会話はすこぶる平凡であった。ただ女が三四郎の筋向こうに腰をかけたばかりである。それで、しばらくのあいだはまた汽車の音だけになってしまう。

次の駅で汽車がとまった時、女はようやく三四郎に名古屋へ着いたら迷惑でも宿屋へ案内してくれと言いだした。一人では気味が悪いからと言って、しきりに頼む。三四郎ももともとだと思った。けれども、そう快く引き受ける気にもならなかった。なにしろ知らない女なんだから、すこぶる躊躇《ちゅうちょ》したにはしたが、断然断る勇気も出なかったので、まあいいかげんな生返事《なまへんじ》をしていた。そのうち汽車は名古屋へ着いた。

大きな行李《こうり》は新橋《しんばし》まで預けてあるから心配はない。三四郎はてごろなズックの鞆《かばん》と傘《かさ》だけ持って改札場を出た。頭には高等学校の夏帽をかぶっている。しかし卒業したしるしに徽章《きしょう》だけはもぎ取ってしまった。昼間見るとそこだけ色が新しい。うしろから女がついて来る。三四郎はこの帽子に対して少々きまりが悪かった。けれどもついて来るのだからしかたがない。女のほうでは、この帽子をむろん、ただのきたない帽子と思っている。

九時半に着くべき汽車が四十分ほど遅れたのだから、もう十時はまわっている。けれども暑い時分だから町はまだ宵《よい》の口のようににぎやかだ。宿屋も目の前に二、三軒ある。ただ三四郎にはちとりっぱすぎるように思われた。そこで電気燈のついている三階作りの前をすまして通り越して、ぶらぶら歩いて行った。むろん不案内の土地だからどこへ出るかわからない。ただ暗い方へ行った。女はなんともいわずについて来る。すると比較的寂しい横町の角《かど》から二軒目に御宿《おんやど》という看板が見えた。これは三四郎にも女にも相応なきたない看板であった。三四郎はちょっと振り返って、一口《ひとくち》女にどうですと相談したが、女は結構だというんで、思いきってずっとはいった。上がり口で二人連れではないと断るはずのところを、いらっしやい、——どうぞお上がり——御案内——梅《うめ》の四番などとのべつにしやべられたので、やむをえず無言のまま二人とも梅の四番へ通されてしまった。

下女が茶を持って来るあいだ二人はほんやり向かい合ってすわっていた。下女が茶を持って来て、お風呂《ふろ》をとと言った時は、もうこの婦人は自分の連れではないと断るだけの勇気が出なかった。そこで手ぬぐいをぶら下げて、お先へと挨拶《あいさつ》をして、風呂場へ出て行った。風呂場は廊下の突き当り便所の隣にあった。薄暗くって、だいぶ不潔のようである。三四郎は着物を脱いで、風呂桶《ふろおけ》の中へ飛び込んで、少し考えた。こいつはやっかいだとじゃぶじゃぶやっていると、廊下に足音がする。たれか便所へはいった様子である。やがて出て来た。手を洗う。それが済んだら、ぎいと風呂場の戸を半分あけた。例の女が入口から、「ちいと流しましょうか」と聞いた。三四郎は大きな声で、

「いえ、たくさんです」と断った。しかし女は出ていかない。かえってはいって来た。そうして帯を解きだした。三四郎といっしょに湯を使う気とみえる。べつに恥かしい様子も見えない。三四郎はたちまち湯槽《ゆおね》を飛び出した。そこそこにかたをふいて座敷へ帰って、座蒲団《ざぶとん》の上にすわって、少なからず驚いていると、下女が宿帳を持って来た。

三四郎は宿帳を取り上げて、福岡県|京都郡《みやこぐん》真崎村《まさきむら》小川《おがわ》三四郎二十三年学生と正直に書いたが、女のところへいってまったく困ってしまった。湯から出るまで待っていればよかったと思ったが、しかたがない。下女がちゃんと控えている。やむをえず同県同郡同村同姓|花《はな》二十三年とでたらめを書いて渡した。そうしてしきりに団扇《うちわ》を使っていた。

やがて女は帰って来た。「どうも、失礼いたしました」と言っている。三四郎は「いいや」と答えた。

三四郎は鞆の中から帳面を取り出して日記をつけた。書く事も何もない。女がいなければ書く事がたくさんあるように思われた。すると女は「ちょいと出てまいります」と言って部屋《へや》を出ていった。三四郎はますます日記が書けなくなった。どこへ行っただろうと考え出した。

そこへ下女が床《とこ》をのべに来る。広い蒲団を一枚しか持って来ないから、床は二つ敷かなくてはいけないと言うと、部屋が狭いとか、蚊帳《かや》が狭いとか言ってちがいがあかない。めんどうがうようにもみえる。しまいにはただいま番頭がちょっと出ましたから、帰ったら聞いて持ってまいりますと言って、頑固《がんこ》に一枚の蒲団を蚊帳いっばいに敷いて出て行った。

それから、しばらくすると女が帰って来た。どうもおそくなりましてと言う。蚊帳の影で何かしているうちに、がらんがらんという音がした。子供にみやげの玩具が鳴ったに違いない。女はやがて風呂敷包みをもとのとおりに結んだとみえる。蚊帳の向こうで「お先へ」と言う声をした。三四郎はただ「はあ」と答えたままで、敷居に尻《しり》を乗せて、団扇を使っていた。いっそこのままで夜を明かしてしまおうかとも思

```
## Step 3 準備したデータを用いてWord2Vecのモデル作成、学習実施
```

```
# Word2Vecライブラリの導入  
#!pip install gensim
```

```
# Word2Vecライブラリのロード  
from gensim.models import word2vec
```

Word2vecを用いて、三四郎の文章中の単語をベクトル化する

```
# size: 圧縮次元数  
# min_count: 出現頻度の低いものをカットする  
# window: 前後の単語を拾う際の窓の広さを決める  
# iter: 機械学習の繰り返し回数(デフォルト:5)十分学習できていないときにこの値を調  
# model.wv.most_similarの結果が1に近いものばかりで、model.dict['wv']のベクトル値  
# ときは、学習回数が少ないと考えられます。  
# その場合、iterの値を大きくして、再度学習を行います。
```

```
# 事前準備したword_listを使ってWord2Vecの学習実施  
model = word2vec.Word2Vec(word_list, size=100,min_count=5,window=5,iter=100)
```

```
# 結果の確認1  
# 一つ一つの単語は100次元のベクトルになっています。  
# 「世間」のベクトル値を確認します。  
print(model.__dict__['wv']['世間'])
```

世間のベクトル

```
# 結果の確認2  
# 関数most_similarを使って「世間」の類似単語を調べます  
ret = model.wv.most_similar(positive=['世間'])  
for item in ret:  
    print(item[0], item[1])
```

0.10964388	-0.980429	0.4997035	-0.4384769	0.20593706	-0.5660683
-0.1424542	-0.08164824	0.00131047	-0.38574845	0.23154311	-0.20020364
-0.7915262	-0.04990723	-0.7188118	-0.24098064	0.15107384	-0.2154366
0.31235328	-0.34722945	-0.20758358	0.4617686	-0.98307246	-0.49646398
-0.10612024	0.16407897	-0.80468506	0.635729	-0.16267775	0.45146376
0.47648802	0.34865445	-0.32653096	-0.00919358	-0.04917296	0.86431086
-0.20292218	-0.14532332	0.27382362	-0.10741992	0.32628235	-0.14972098
-0.166223	0.72024727	0.07788672	0.20204976	-0.20483725	0.05324753
-1.1225064	0.19487472	0.38060167	-0.26776007	0.05262767	-0.31786847
0.39386213	-0.11356686	0.07396717	-0.717772	0.73588693	-0.12168456
0.00547593	0.15277809	-0.4571755	0.21926825	-0.653149	-0.05090689
-0.25308615	0.42054158	0.07740451	-1.0023556	-0.11924236	-0.49817115
0.04489631	0.02224448	0.5508423	0.89573216	0.24378105	0.30015174
0.393421	-0.04692438	0.02466744	-0.13444859	-0.07020038	-0.79955477
0.14002748	0.10524069	-0.3855506	0.38433182	0.11913352	0.09627419
-0.8063719	-0.97816753	-0.34846506	0.6854964	0.21095207	0.00897974
0.20270449	0.6240339	-0.37366846	-0.28756556		

世間に類似した単語

聞こえる	0.5731331706047058
自己	0.5621089339256287
喝采	0.5495569109916687
決心	0.5201483964920044
今日	0.5122823715209961
賛成	0.49857455492019653
社会	0.49650177359580994
酔う	0.49364516139030457
堪える	0.48814666271209717
喜ぶ	0.48021504282951355